AI Infrastructure Reference Architecture

IBM Systems

87016787USEN-00

Authors

Kelvin Lui, kelvinl@ca.ibm.com

Jeff Karmiol, jkarmiol@ca.ibm.com

Table of Contents

1. Introduction

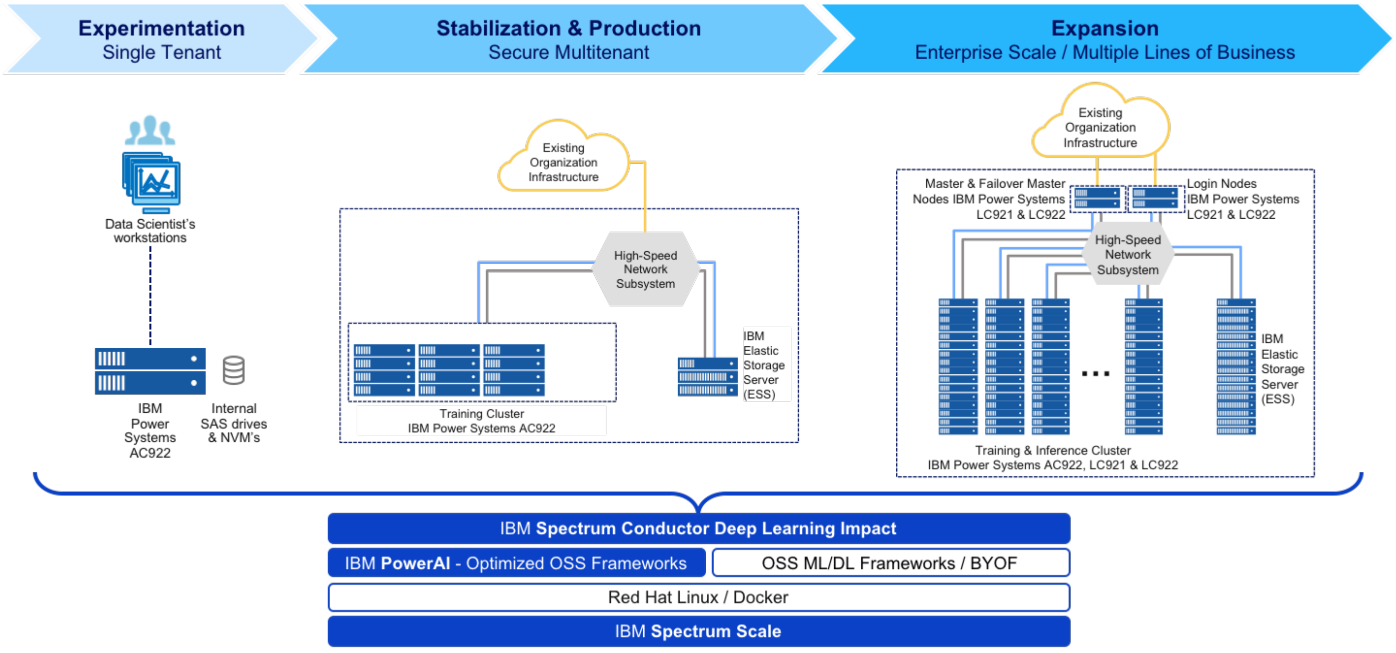
Organizations are using Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) to develop powerful new analytic capabilities spanning multiple usage patterns, from computer vision and object detection, to improved human computer interaction through natural language processing (NLP), to sophisticated anomaly detection capabilities. At the heart of any use case associated with AI/ML/DL are sophisticated pattern recognition and classification capabilities which serve as the birthplace for revolutionary applications and insights of the future.

IBM makes AI/ML/DL more accessible and more performant. By providing a production proven AI environment made up of the following components, organizations can rapidly deploy a fully optimized and supported platform for AI with blazing performance.

- IBM **PowerAI Enterprise** software platform
- IBM **Power Systems** servers
- IBM **Spectrum Scale on all-flash Elastic Storage Server**

1.1. Purpose

This document is intended to be used as a reference by data scientists and IT professionals who are defining, deploying and integrating AI solutions into an organization. It describes an architecture that will support a productive proof of concept (PoC), experimental application, and sustain growth into production as a multitenant system that can continue to scale to serve a larger organization, while integrating into the organization's existing IT infrastructure.


Experimentation
Single Tenant
Stabilization & Production
Secure Multitenant
Expansion
Enterprise Scale / Multiple Lines of Business
Data Scientist's workstations
IBM Power Systems AC922
Internal SAS drives & NVM's
Existing Organization Infrastructure
High-Speed Network Subsystem
IBM Elastic Storage Server (ESS)
Training Cluster IBM Power Systems AC922
Existing Organization Infrastructure
Master & Failover Master Nodes IBM Power Systems LC921 & LC922
Login Nodes IBM Power Systems LC921 & LC922
High-Speed Network Subsystem
IBM Elastic Storage Server (ESS)
Training & Inference Cluster IBM Power Systems AC922, LC921 & LC922
IBM Spectrum Conductor Deep Learning Impact
IBM PowerAI - Optimized OSS Frameworks
OSS ML/DL Frameworks / BYOF
Red Hat Linux / Docker
IBM Spectrum Scale

One software stack from experimentation to production

1.2. Target Audience

The target audience for this document includes

- **Data scientists** who are using or consider using AI models to derive value from data

- **Administrators** who are tasked with providing AI environments to an organization and responsible for keeping them running
- **Architects** who need to think about how to integrate AI frameworks into their existing infrastructure and how to scale to serve an organization
- **IT Departments** who will are responsible for the service infrastructure to support these initiatives

1.3. Terminology

- **AI** – When the abbreviation AI is used in this reference document it refers to the larger topic that includes Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL).
- **GPU** – Graphical Processing Units which have expanded from their prior role as a graphics accelerator in to a highly dense parallel computing engine.
- **HDFS** – Hadoop Data File System is a common scale-out file system using storage rich servers for analytics and machine learning.
- **PoC** – Proof of Concept, who's purpose is to demonstrate the ability of a system to perform an activity, usually against a defined set of criteria. In this case, a PoC would be to demonstrate that a solution based on this reference architecture delivers the benefits and values claimed.

2. Requirements for AI

- Accelerated server infrastructure: today the dense computing power of GPUs are required to make deep learning possible
- Storage is one of the first most overlooked aspects of AI initiatives. Data scientists initially focus on the frameworks and models and only encounter storage issues when they begin to scale.
- Data sources. Initial connections and ongoing managements
- Serving multiple users, groups, lines of business and organizations
- Centralizing compute (i.e., GPU) and storage resources and sharing with multiple users, including consolidating disparate infrastructure while ensuring QoS for stakeholders
- Policies for sharing resources like fairshare, preemption and the ability borrow and reclaim resources
- Security **-** Most of the AI frameworks, toolkits and applications available today do not implement security at all, relegating them to disconnected experiments and lab implementations.
- Fault tolerance and availability
- Performance by distributing the workloads (e.g., data ingest, preparation and model training) across multiple GPUs and servers
- Data preparation environments including Spark and Hadoop, and tools to change data formats and adjust sizes
- Removing limitations in current AI frameworks and model implementations
  - Fixed and hardcoded GPU and network topology maps
  - Requiring all GPU resource to be available before training can start
  - Flexible and changing GPU allocations while jobs are executing, allowing GPUs to be added to and removed from a training job without the need to stop and restart
- Resiliency to GPU, server and network failures or outages

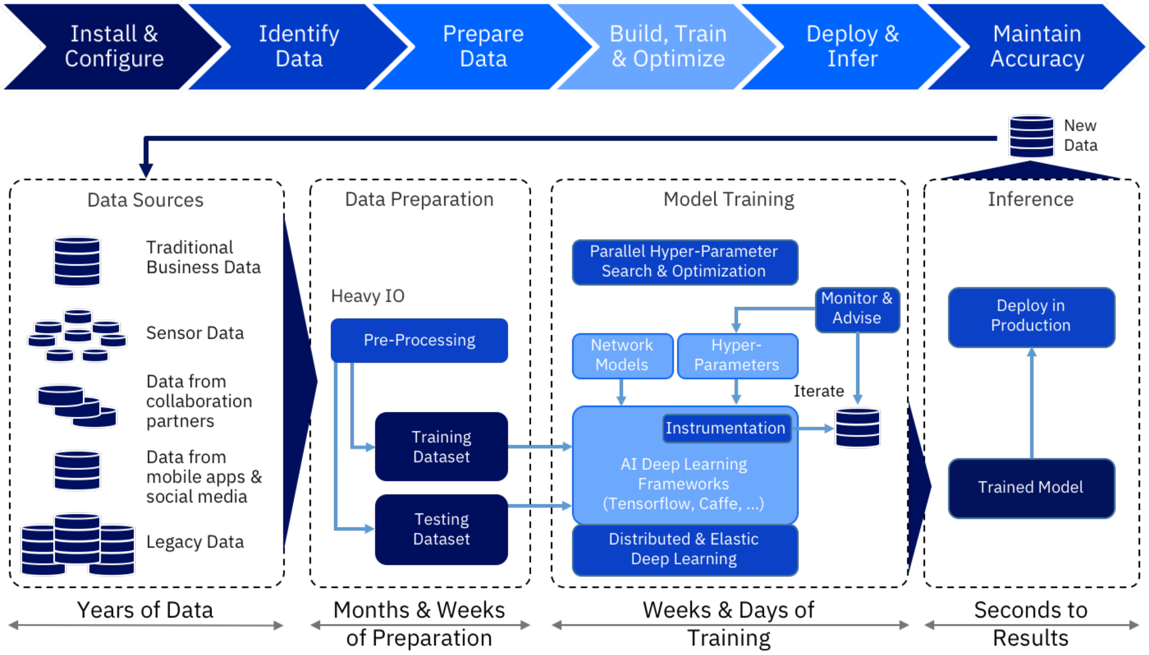
- Support gaps in data science knowledge like selecting and optimizing hyperparameters
- Help reduce wasted time by reducing iterations and unnecessary tasks
- Simplify models testing and inferencing
- Support all open source AI frameworks
- Custer management and monitoring

3. Concepts

AI projects are iterative multi-stage data-driven processes or workflows and require specialized knowledge, skills and usually new compute and storage infrastructure. These projects have many attributes that are familiar to traditional CIOs and IT departments. The first of which is that the results are only as good as the data going into it, and model development is dependent upon having a lot of data and the data being in the format expected by the deep learning framework. It is also iterative; repeatedly looping through data sets and tunings to develop accurate models, then comparing new data in the model to the original business or technical requirements to refine the approach.

3.1. AI Workflow

The AI workflow is really a process cycle, once you produce a trained neural network model you go back and retrain the model with new data to keep it current and to improve its accuracy. Some organizations are using the champion challenger approach to achieve the most effective results by comparing an existing model, champion, to new models being developed, challengers. After the results are compared, the model with the best results becomes the new champion.



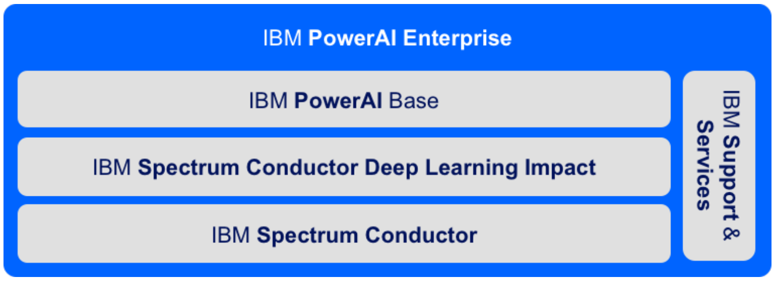
3.1.1. Design & Grow

AI is a sophisticated or complex process that requires specialized software and infrastructure. AI projects are notoriously hard to get up and running from installing the systems and software, to building the model frameworks with all the right libraries and drivers, then finding and preparing the data. Most organizations start with small pilot projects bound to a few systems and data sets.

As projects grow beyond the first test systems, the appropriate storage and networking infrastructure is needed so the AI infrastructure can sustain growth and eventually support an organization. The design goal for this reference architecture is to dramatically reduce the complexity and time required to plan growth and broader adoption. It is also critical to build a system that connects to and integrates into the existing organizational infrastructure.

Choosing a supportable platform with the right integrated software from the beginning (i.e., at the PoC/experimentation stage) will lower barriers to growth and broader adoption. For example, the IBM **Power Server AC922** and **S822LC for HPC** both come equipped with NVIDA GPUs that are connected to the CPU via NVLink technology.

IBM **PowerAI Enterprise** and IBM **Spectrum Scale** delivers the scalable software stack presented in this reference architecture. They are designed for ease of installation, reduced time to first use and supports the single system PoC and is designed to scale to thousands of systems. By downloading a few packages, the full software stack can be installed, configured and ready to use in a matter of hours. IBM PowerAI Enterprise includes full versions of IBM **PowerAI** base, IBM **Spectrum Conductor** and IBM **Spectrum Conductor Deep Learning Impact**. IBM Spectrum Scale is software-defined storage that can run on practically any storage, and powers the IBM **Elastic Storage Server** (ESS).

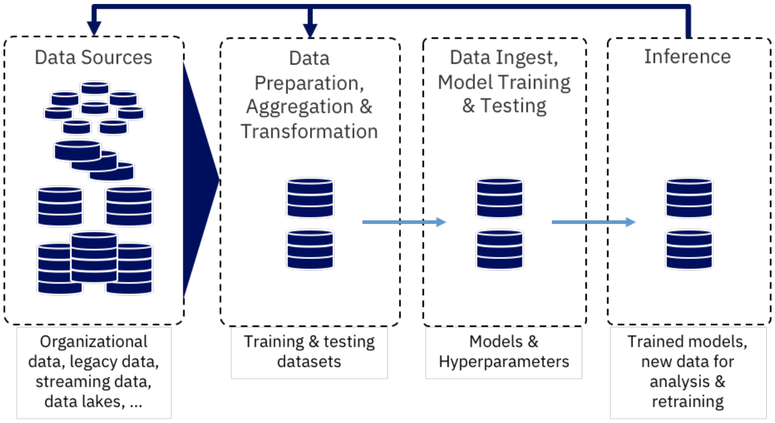


3.1.2. Data Preparation and Transformation

The accuracy and quality of a trained AI model are directly affected by the quality and quantity of data used for training. The data scientist needs to understand the problem they are trying to solve and then find the data needed to build a model to solve the problem.

Data in the context of AI is separated into a few broad sets, the data used to train and test the models, the new data that is analyzed by the models and the historical or archived data that may be reused. This data can come from many different sources such as traditional organizational data from ERP systems, databases, data lakes, sensors, collaborators and partners, public data, mobile apps, social media, and legacy data, in may be structured and unstructured in many formats such as file, block, object and Hadoop Distributed File Systems (HDFS).

Many AI projects begin as a big data problem. Regardless of the source, a large volume of data is needed, and it inevitably needs preparation, transformation and manipulation.


Data Sources
Data Preparation, Aggregation & Transformation
Data Ingest, Model Training & Testing
Inference
Organizational data, legacy data, streaming data, data lakes, ...
Training & testing datasets
Models & Hyperparameters
Trained models, new data for analysis & retraining

The AI models require the training data to be in a specific format, each model has their own and usually different format, and invariably the data is nowhere near those formats. Preparing the data is often one of the largest organizational challenges, not only in complexity, but also in the amount of time it takes to transform the data in to a format which can be analyzed. Many data scientists claim that over 80% of their time is spent in this phase and only 20% on the actual art of data science. Data transformation and preparation is typically a highly manual and serial set of steps: identifying and connecting to data sources, extracting to a staging server, tagging the data, using tools and scripts to manipulate the data (e.g., removing extraneous elements, breaking large images down to 'tile' size so they will fit in GPU memory, etc.) Hadoop is often a significant source of this raw data, and Spark is typically the analytics and transformation engines used along with advanced AI data matching and traditional SQL scripts.

There are two considerations in this phase. One is the data storage and access and the other is the speed of execution. To speed up this stage, this reference architecture uses IBM Spectrum Scale, which provides multi-protocol support with a native HDFS connector, to centralize and analyze data in place, rather than wasting time copying and moving data.

To accelerate the data processing, the IBM Spectrum Conductor component of IBM PowerAI Enterprise provides a production Apache Spark distribution and a distributed Spark and application environment to help automate and run these tasks in parallel. IBM Spectrum Conductor is also used to establish and maintain connections to storage resources and to capture data formatting information, enabling faster iterations through these time-consuming tasks.

3.1.3. Build, train, optimize models

This stage of the AI process is compute heavy and very iterative. There is both art and science behind this process, and it is at this stage that an organization's gap in data science skills hurts the most. The model hyperparameters need to be set. They are the configuration values used to define the characteristics of the models, such as the number of leaves and depth of a tree, the number of hidden layers, and other factors, need to be chosen before starting the training and this requires very specialized knowledge, many iterations and a lot of trial and error. Also, the open source frameworks are rigid and fragile, and once the training begins the number of GPUs cannot change and if there is any change, like a GPU or network failure or network outage, will cause the job to fail. This becomes a problem if a model that takes multiple days to train fails a couple of hours before its schedule to finish, in this case the job needs to be rerun from the beginning, wasting all the elapsed training time.

This reference architecture helps to reduce the data science knowledge gap by using cognitive algorithms in the IBM Spectrum Conductor Deep Learning Impact component of IBM PowerAI Enterprise, to suggest and optimize hyperparameters. Failures are avoided, and resource sharing is enabled by using **elastic distributed training** capabilities. Resource (i.e., GPUs) are dynamically allocated at runtime, allowing models to train on multiple GPUs across multiple systems, and the allocation is flexible during runtime, so GPUs can be added or removed from a training job without the need to kill the job and start over. This elastic training capability enables resource sharing, job pre-emption and priority.

To help assist in determining the quality of the hyperparameters and the accuracy of the model, **runtime training visualization** allows the data scientist to see the progress of model while the training is executing. Iteration, loss, accuracy and histograms of weights, activations, gradients of the neural network are presented, which provides the opportunity to stop training if the training is not returning the right or accurate results. This helps eliminate wasted time by stopping long running training that are not being productive.

3.1.4. Deploy, infer, score model and capture organizational value

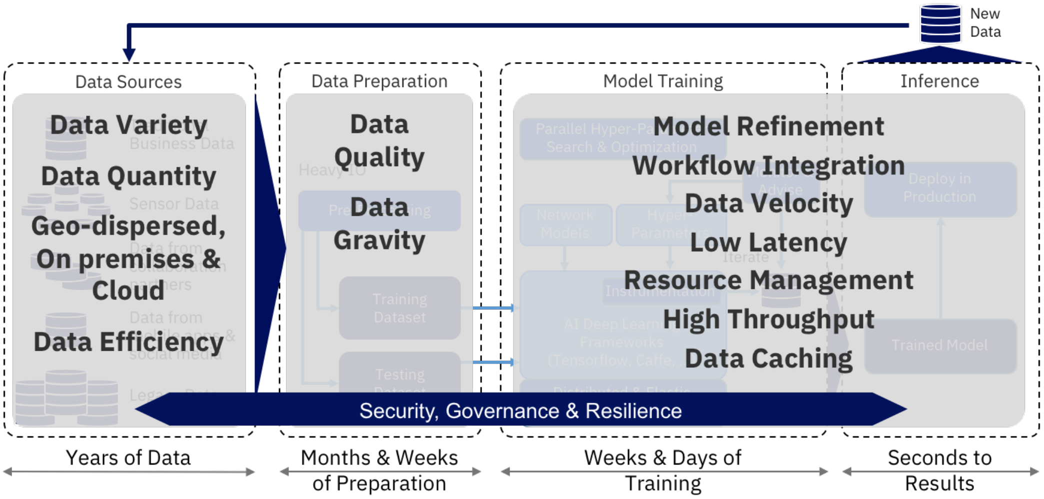
AI models are trained initially on historical data and are then tested and deployed and run as an inference service and used to analyse new and real time data. This stage of the process is where organizations realize the value from the AI project.

IBM **Spectrum Conductor Deep Learning Impact** software component provides an interface where with the single click the AI model can be deployed as an inference service running on the shared training cluster, enabling inference requests to be submitted through a browser and a REST API.

3.1.5. Maintain model accuracy, ingest and train on new data

Ensuring a deployed model remains up to date and gets smarter as additional data is gathered make this a cyclical process. Completing the AI cycle, the newly analyzed data is fed back into the process to retrain and improve the model or used to build new models. The data flow is circular and requires large volumes of storage to handle the large datasets through the many steps along the AI workflow. The IBM Spectrum Conductor Deep Learning Impact UI retains all the data connections and setting (business logic) from the previous training runs, simplifying subsequent runs.

3.2. The AI Data Pipeline



Different stages of the AI pipeline have different data and processing requirements. All data within an organization is fair game for AI initiatives including new data streaming in from the edge of the organization, current structured business data to archive data stored in a data lake. For the data ingest, preparation and training stage, there may be years of data to process and it can take weeks or even months to complete. By contrast with these extended time frames, once deployed, the inference system may need to respond in seconds.

Storage performance is critical for workflow throughput. Although many of the training sets are cached at the server, each GPU wants at least a GB/s of data throughput to keep it fed when larger data sets are used, or when switching data sets.

Often overlooked in the PoC stage, but quickly becomes an issue when scaling to production, storage capacity is an important consideration during the training stage. The data may also be in different formats in different systems, so multi-protocol capability may be needed. The data may also be geographically dispersed, an additional factor the storage system needs to handle. Once deployed for inference, fast access to the data becomes particularly important to support the response requirements of users and applications, which typically need answers in seconds.

A system such as IBM Spectrum Scale and implemented as IBM Elastic Storage Server (ESS) is perfectly suited to meeting these requirements. It is a high-performance system running on IBM Power Server systems that can scale out to handle petabytes or exabytes of data. It supports a wide variety of protocols for accessing file or object. For Hadoop applications it provides direct access to data without having to copy the data to HDFS, as is usually required. Avoiding the overhead of copying data between systems lowers cost by saving space and speeds time to results.

4. Architecture Overview

4.1. Frameworks and Models

AI frameworks provide the building blocks for data scientists and developers to design, train and validate AI models through a high-level programming interface and without getting into the nitty-gritty of the underlying algorithms.

AI models are the work products produced by working with the AI frameworks, these are the objects that get trained using data and then used to infer (i.e., perform analysis) on new data.

TensorFlow and Caffe are a couple of the most popular frameworks being used today, and there are more models available, and more being developed all the time as AI technology matures. The IBM PowerAI base component of IBM PowerAI Enterprise delivers open source frameworks that are ready to use (i.e., complied with all required libraries and drives), optimized, tested and supported from IBM. The software is easily downloaded and installed and ready to use in a few hours.

In addition to the frameworks supplied by IBM PowerAI the data scientists can run almost any framework on the system, this capability is informally called BYOF, Bring Your Own Framework.

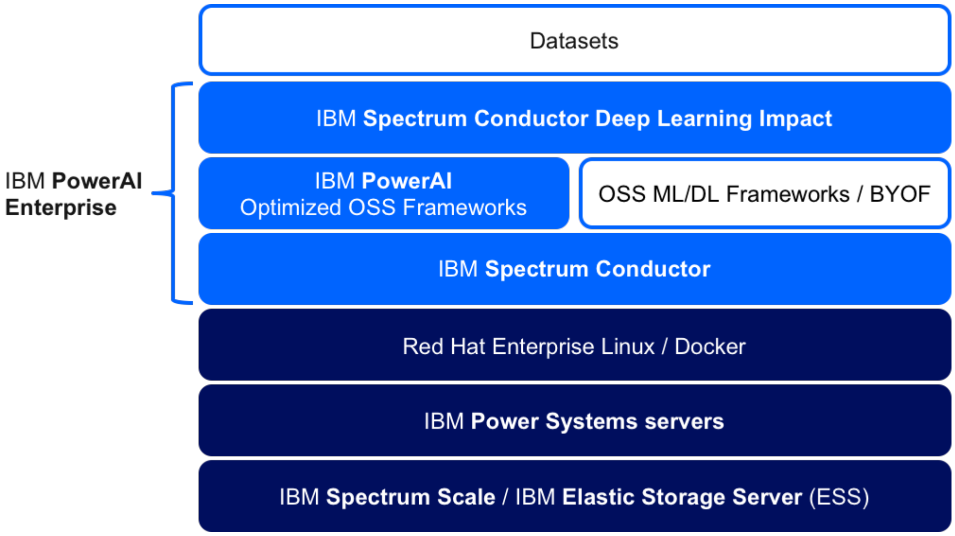
4.2. Platform

This reference architecture describes software components, the software stack, used to build the AI environment as the *platform*. This includes the AI frameworks, the software used to build a multitenant data science environment, the distributed computing, training and inference environment, and a tiered data management environment.

One of the goals of this reference architecture is to provide a *single* software stack that will simplify a PoC and scale to thousands of nodes. As well as addresses some of the bigger challenges facing developers and data scientists — significantly cutting down the time required for AI system training and simplifying the development experience. This reveals itself in a number of value differentiators for AI applications that IBM **PowerAI Enterprise** brings to the table from IBM **PowerAI**, IBM **Spectrum Conductor**, IBM **Spectrum Conductor Deep Learning Impact** and IBM **Spectrum Scale**.

- IBM **PowerAI** is a software distribution package containing many of the major open source deep learning (DL) frameworks for model training, such as TensorFlow and Caffe, and their associated libraries. These frameworks complied with all necessary drivers and libraries, optimized for performance by using the NVLink-based IBM Power Systems servers and ready to use as soon as they are installed, https://www.ibm.com/us-en/marketplace/deep-learning-platform/details.
- IBM **Spectrum Conductor** is a highly available multitenant application designed to build a shared, enterprise-class distributed environment for deploying and managing modern computing frameworks and services, such as Spark, Anaconda, TensorFlow, Caffe, MongoDB and Cassandra. Spectrum Conductor also provides centralized management and monitoring, along with end-to-end security, https://www.ibm.com/ca-en/marketplace/spark-workload-management.

- IBM **Spectrum Conductor Deep Learning Impact** is a software component addition to IBM Spectrum Conductor that builds a deep learning environment providing an end-to-end workflow that allows data scientists to focus on training, tuning and deploying models into production, https://www.ibm.com/ca-en/marketplace/spectrum-conductor-deep-learning-impact, https://www.ibm.com/ca-en/marketplace/spectrum-deep-learning-impact.
- IBM **Spectrum Scale** is an enterprise-grade parallel file system that provides superior resiliency, scalability and control. IBM Spectrum Scale delivers scalable capacity and performance to handle demanding data analytics, content repositories and technical computing workloads. Storage administrators can combine flash, disk, cloud, and tape storage from all across the organization into a unified system with higher performance and lower cost than traditional approaches, https://www.ibm.com/ca-en/marketplace/scale-out-file-and-object-storage.


Datasets
IBM PowerAI Enterprise
IBM Spectrum Conductor Deep Learning Impact
IBM PowerAI Optimized OSS Frameworks
OSS ML/DL Frameworks / BYOF
IBM Spectrum Conductor
Red Hat Enterprise Linux / Docker
IBM Power Systems servers
IBM Spectrum Scale / IBM Elastic Storage Server (ESS)

4.3. Infrastructure

This reference architecture describes an infrastructure on which is this AI solution runs, and includes the compute servers, high-performance storage and high-speed networking.

4.3.1. Compute

The resource of primary interest for AI is the GPUs, and they are located on the IBM Power Systems server nodes. They feature CPU to GPU NVLink connection, which delivers much higher I/O bandwidth than x86 based servers, and are also capable of supporting large amounts of systems memory.

- IBM **Power System AC922** feature POWER9 CPUs and support 2-6 NVIDIA Tesla V100 GPUs with NVLink providing CPU:GPU bandwidth speeds of 100 GB/sec air cooled or 150 GB/sec water cooled. This system supports up to 2TB total memory, https://www.ibm.com/id-en/marketplace/power-systems-ac922.
- IBM **Power System S822LC for HPC** feature POWER8 CPUs and support 2-4 NVIDIA Tesla P100 GPUs with NVLink GPUs providing CPU:GPU bandwidth speeds of 64 GBps. This

system supports up to 1TB total memory, https://www.ibm.com/id-en/marketplace/high-performance-computing.

4.3.2. Storage

To support the variety and velocity of data used and produced by AI, the storage system needs to be intelligent, have enough capacity and be high-performance. Key attributes include tiering and public cloud access, multi-protocol support, security and extensible metadata to facilitate data classification. Performance is multi-dimensional for data acquisition, preparation and manipulation, high-throughput model training on GPUs and latency sensitive inference.

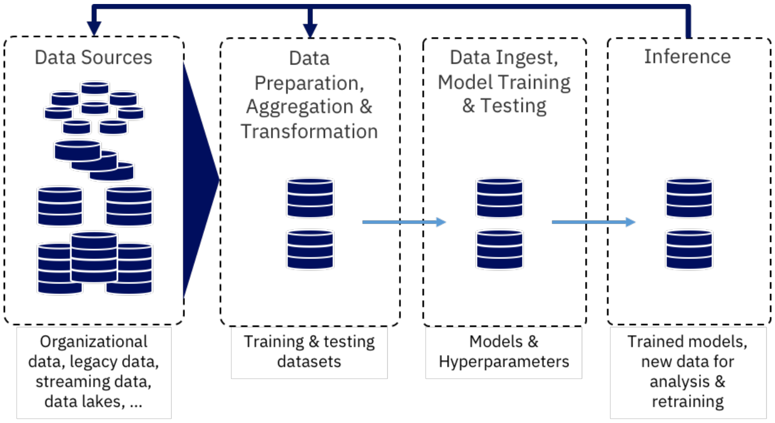
The type of storage used depends on the location of the data and the stage of processing for AI. Corporate and legacy data will usually reside in an organizational data lake in HDFS.

- IBM **Elastic Storage Server** (ESS) combines IBM **Spectrum Scale** software with IBM POWER8 processor-based I/O-intensive servers and dual-ported storage enclosures. IBM Spectrum Scale is the parallel file system at the heart of IBM ESS and scales system throughput as it grows while still providing a single namespace. This eliminates data silos, simplifies storage management and delivers high performance. By consolidating storage requirements across your organization onto IBM ESS, you can reduce inefficiency, lower acquisition costs and support demanding workloads, https://www.ibm.com/id-en/marketplace/ibm-elastic-storage-server.
- IBM **Spectrum Scale** can also be deployed on the next generation IBM NVMe all-flash arrays, which will be available in mid-2019. This solution will make use of the latest in NVME FlashCore Modules in a dense package. In internal IBM lab pre-release testing, this configuration delivered 40 GB/s in 1M random read performance in a single IBM NVMe array.

4.3.3. Networking

The exact architecture, vendors and components needed to build the network subsystem depend upon the organizational preference and skills. InfiniBand and high-speed Ethernet are a requirement, as is a network topology that allows both north/south (server to storage) traffic and can also support east/west (server to server) traffic. Adopting a topology that extends to an InfiniBand Island structure will allow the training environment scale for large clusters.

Keep in mind that an adequate network subsystem with necessary throughput and bandwidth to connect the different tiers of storage together, whatever from the edge into the primary storage and data lake, and what’s needed from primary storage and data lake into the data prep staging tier and then into the training cluster where performance and throughput is the primary requirement needed to feed the data hungry GPUs for training.


Data Sources
Data Preparation, Aggregation & Transformation
Data Ingest, Model Training & Testing
Inference
Organizational data, legacy data, streaming data, data lakes, ...
Training & testing datasets
Models & Hyperparameters
Trained models, new data for analysis & retraining

High bandwidth and low latency between the storage and compute nodes is absolutely critical, and sufficient bandwidth between the nodes needs to also be considered for data ingest and transformation phase of the workflow. Performance is key when training models to make sure sufficient data is delivered to the systems to keep the GPUs running at capacity, so a high-speed network subsystem is needed for the training cluster (i.e., Fast ethernet and InfiniBand).

5. Enterprise Readiness

This reference architecture describes a system that is suitable for and has been deployment in enterprise environments including regulated environments, such as top banks in the US, Canada and China. This enterprise suitability applies to the physical infrastructure (i.e., server, storage and network solutions), and the software platform (i.e., OS, storage management, workflow, resource and workload management solutions).

5.1. Return on Investment (ROI)

AI projects requires a new approach to infrastructure especially for deep learning, this reference architecture presents a solution that will maximize your investment to get the highest utilization (i.e., the most use) out of your infrastructure investment.

- The GPU accelerated IBM **Power Systems AC922** and **S822LC** servers specified in this reference architecture can deliver 2 to almost 4 times higher throughput and performance than commodity servers currently available. This means you can do more work with less servers, requiring the purchase of less servers to do the same or more work. See the performance proof-points at https://developer.ibm.com/linuxonpower/perfcol/perfcol-mldl/.
- The IBM **Spectrum Conductor Deep Learning Impact** software component maximizes the utilization of your server and especially GPU resources with granular and dynamic workload allocation and Elastic Distributed Training. The intelligent scheduler allocates resources to the workload faster and more efficiently than other workload and resource managers (https://www.ibm.com/....<<url>> to latest STAC report.), making sure that no compute and GPU cycles are left idle as long as there is workload to process. This means that you get the maximum out of your server and GPU resources and only need to add more resources as workload demands and not due to unknown and wasted idle resources. Elastic Distributed Training dynamically assigns GPU resources to the training job at

execution time, and also allows for GPUs to be added and removed from a job without the need to stop the job and starting over. Elastic Distributed Training separates resource allocation from the AI model eliminating the need to hardcode the network and GPU topology, simplifying the development and ongoing training efforts.

- The multitenant capabilities of the IBM Spectrum Conductor software component allow you to consolidate multiple compute cluster silos, eliminating wasted idle time, reducing administrative overhead and providing an overall cluster with larger capacity. The larger resource capacity along with high utilization maximizes current resources and delays the need to acquire additional resources. The resource policies enable Quality of Service (QoS) where an amount of resources can be guaranteed to individual tenants while also allowing for sharing and graceful return as workload permits. Prioritization and graceful preemption is also available making sure the right workloads are run at the right time. Together the resource policies provide you with the ability to manage significant workload variability, throughput and performance across multiple users and applications.
- The computing and storage management platforms specified in this reference architecture are heterogeneous and can take advantage and manage existing resources in your environment, extending the return on your older investment and reducing the overall administrative overhead. IBM Spectrum Conductor enables existing non-accelerated POWER8 and POWER9 based servers and x86 based servers to be managed and included into your AI environment.
- In addition to managing the IBM ESS specified in this reference architecture, IBM Spectrum Scale can also be used to manage existing non-IBM storage systems in your environment. Unifying storage across the organization, eliminating the need to copy into HDFS and tiering data to the appropriate storage level all contribute to maximizing return on investment for storage.

5.2. Operating System

As part of the development of PowerAI and PowerAI Enterprise, IBM has integrated multiple open source components to RedHat Enterprise Linux (RHEL) little endian in order to support deep learning at production scale. RHEL is considered to be an “enterprise class” operating system (OS) by most organizations because it provides an open, reliable and scalable foundation to tackle challenging AI workloads and is backed by IBM support and services.

5.3. Reliability and Availability

Reliability and availability are cornerstones for deploying systems into an enterprise environment. The AI function will transition from development to mission critical and when the business begins to rely on the function reliability and availability become critical. The technologies used in this reference architecture have been validated and tested in production systems for decades.

- The IBM Spectrum Conductor and IBM Spectrum Conductor Deep Learning Impact software components provide:
  - Failover master nodes to make sure that IBM Spectrum Conductor remains available in case of the master node becoming unavailable.

  - Elasticity, distributed computing and dynamic resource and workload management across multiple systems ensure that the location and state of all jobs are tracked and that they are run to completion or exit. If a resource, GPU or node, becomes unavailable then the jobs that were running in that resource will be rerun on another available resource. This behavior repeats itself until no more resources are available and creates a highly available system.
- IBM Elastic Storage Server (ESS) and IBM Spectrum Scale, is a highly reliable and redundant storage solution. Data is efficiently distributed with erasure coding across disks.
- The IBM Power Systems AC922 and S822LC for HPC servers employ the following reliability, availability and serviceability features: Processor instruction retry, Selective dynamic firmware updates, Chip kill memory, ECC L2 cache, L3 cache, Service processor with fault monitoring, Hot-swappable disk bays and Redundant cooling fans.

5.4. Security

Most of the AI frameworks, toolkits and applications available do not implement security, relegating them to disconnected experiments and lab implementations. The IBM Spectrum Conductor software component and IBM Spectrum Scale wrap security features around these frameworks allow for production deployments into many regulated organizations including major US, Canadian, European and Chinese financial and governmental institutions. To satisfy today's need for security and privacy, these solutions employ the latest security protocols and have been subjected to extensive security scanning and penetration testing.

The IBM Spectrum Conductor software component implements end-to-end security, from data acquisition and preparation to training and inference. Security is implemented around:

- **Authentication** - Support for Kerberos, Siteminder, AD/LDAP and OS authentication, including Kerberos authentication for HDFS.
- **Authorization** - Fine grained access control, ACL/role-based control (RBAC), Spark binary life cycle, notebook updates, deployments, resource plan, reporting, monitoring, log retrieval and execution
- **Impersonation** - Allow different tenants to define production execution users
- **Encryption** - SSL & authentication between all daemons and Storage encryption with IBM Spectrum Scale.

5.5. Planning your Design to Scale

The approach presented in this reference architecture is to start small, with a couple of servers and moderate storage or existing storage, and then scale by adding compute and GPU or storage as needed. The technologies used in this reference architecture are designed to do just that, scale as needed. When additional compute and GPU resources are needed additional IBM Power Systems servers can be added to the environment and without the need for any downtime. The IBM Spectrum Conductor software component and IBM Spectrum Scale both can add additional resources while the systems are running in production.

- The IBM Spectrum Conductor software component, the foundation of the reference architecture software stack, can scale to 4000 servers/nodes or 8000 sockets, where each CPU and GPU count as a single socket

- IBM Spectrum Scale, the storage management system, can scale to manage up to 8 Exabytes (maximum system size) across 16,384 nodes.

5.6. Support and Maintenance

All the components specified in this reference architecture and acquired from IBM (i.e., servers, networking, storage, OS, software platform including AI frameworks) are fully backed by IBM support and services. This includes the open source AI libraries included in IBM PowerAI. In the event that support is required, the same single point of contact is used for all the components sourced from IBM and all cases are owned and managed by IBM.

IBM also manages all software patches and upgrades to the systems, including: TensorFlow, BVLC Caffe, IBM Caffe, Python3 support for TensorFlow and Caffe, Anaconda, PyTorch (Tech Preview), Snap.ml (Tech Preview), NCCL2

5.7. Cloud-ready

The IBM Spectrum Conductor software component and IBM Spectrum Scale are both cloud ready and can be used on-premises, in the cloud (private and public) and in a hybrid cloud configuration. When additional resources are needed to handle peak loads or if the data is already located in the cloud, additional cloud-based compute and storage resources can be easily accessed.

IBM Spectrum Conductor provides cloud bursting feature called Host Factory, that enables your cluster to dynamically burst workloads to cloud hosts. When resource demand exceeds the capacity of the cluster, additional cloud hosts are provisioned and added to the cluster. When there is excess capacity cloud hosts are returned to the cloud providers, providing cost savings, flexibility and scalability. Host factory currently supports IBM Cloud and Amazon Web Services (AWS).

5.8. Greater Efficiency for the Data Scientist

The combination of all the benefits delivers greater efficiency for the data scientist.

- Better performing servers allows for faster training times enabling more iterations, or the same training times with more data
- Easy to install software enables the data scientist to begin their AI project in hours
- Ready to use open source AI frameworks (compiled with all libraries and necessary drivers) eliminates what is reported to be a task that could takes days or weeks
- Multitenancy allows resources to be pooled so the data scientist has access to a more resources
- Elastic distributed training means the data scientist doesn’t need to stand around and wait for resources to be released in order to start training a model
- Hyperparameter search and optimization bridges the data science knowledge gap and reduces the trial and error time needed to get the right values
- Visualization while training allows helps to eliminate time wasted waiting for inaccurate or poorly performing models to complete training
- Unified file system that eliminates the need to copy data into the AI system before it can be used eliminates the long wait times such copying requires before the data can be used

6. Reference Architectures

This reference architecture has been designed to supoport the AI adoption cycle, from experimenting with one or two servers to scaling up the implementation in a way that it can grow to serve multiple tenants and an organization.

- **Proof of Concept / Experimentation**: Data Science Proof of Concept (PoC)
  - One or two accelerated compute nodes
  - Small scale data sets using local or available shared storage
  - Framework selection and prototyping, hyperparameter optimization and inference testing
- **Stabilization and Production**: Adoption of best practices to facilitate continued growth while sharing resources
  - Multiple nodes - Initial cluster deployment, four, eight, twelve, …, accelerated compute nodes
  - Often overlooked, data storage delivering sufficient capacity to handle multiple projects and throughput to support GPUs (local SSD or NVM)
  - Dedicated or high-speed network subsystem sufficient to reduce latency. Thoughts about a network topology that will allow for scale (e.g., InfiniBand islands)
  - Secure multitenant environment providing access to multiple data scientists supporting multiple different frameworks and models
  - Workload and workflow management to speed development and ease transition to production
- **Scale-out and Expansion**: Growing to meet the needs of an organization
  - Scale-out compute and storage cluster infrastructure while providing availability and protection (i.e., no downtime and no data loss)
  - Addition of non-accelerated nodes for ML and Inference as well as service nodes to offload the scheduling master and login node functions
  - Storage support across the data pipeline from ingest to inference
  - Integrated active archive and data tiering
  - Metadata management
  - Inference Services

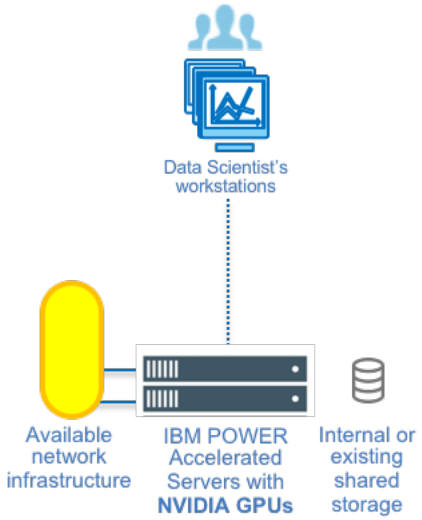
6.1. Proof of Concept Configuration

The Proof of Concept (PoC) is an experiment, proving to the data scientist that they can in fact build an AI environment, deliver trained models and that the solution presented by this reference architecture delivers value. The presented software platform/stack will ease the PoC and also serve as the software stack for the production and scale out environments. The PoC is usually the domain of a single data scientist looking for tools to make help reduce the complexity of the AI process or thinking about how to start scaling their AI projects. Since this is most likely an isolated system, security is usually not an issue.

Anecdotally, 99% of AI model training is performed on a single server, opening this up to a couple of servers and thinking about growth is a step toward institutionalizing AI. Evaluation criteria should involve:

- Installation, up and running experience

- Performance delivered by the IBM Power Systems servers
- Productivity improvements using the IBM Spectrum Conductor Deep Learning Impact software component
- Simpler experience iterating through the AI workflow
- Initial experimentation with multitenancy, perhaps supporting multiple versions of the same framework or application on shared infrastructure


Data Scientist's workstations
Available network infrastructure
IBM POWER Accelerated Servers with NVIDIA GPUs
Internal or existing shared storage

PoC Configuration
IBM Recommended Infrastructure Systems
• IBM Power System AC922 (P9) and/or • IBM POWER S822LC (P8) for HPC
IBM Recommended Solution Software
• IBM PowerAI Enterprise • Red Hat Enterprise Linux for IBM Power Little Endian

6.2. Stabilization and Production

While continuing to use the same software platform/stack as used in the PoC, the training cluster starts to grow and take form. For those who have a background in high performance computing (HPC) the architecture looks like a typical HPC data center.

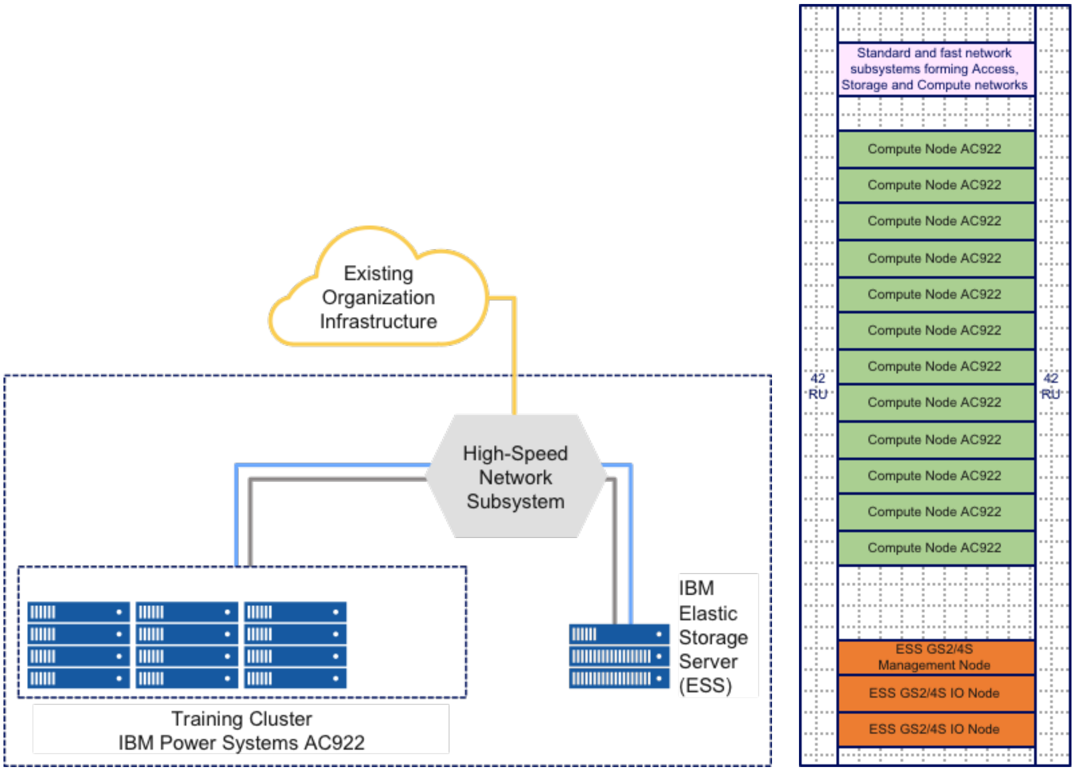
As the implementation begins to form and start **scaling** the PoC into a production system, additional compute servers and dedicated storage will be added. This reference architecture presents compute and storage solutions that allow for either to scale independently as the characteristics and requirements of your environment start to become understood.

Adding dedicated **storage** needs to consider the capacity needed to contain the amount of data that will form the training and testing data sets. This is usually one of the main issues encountered as AI projects begin to move out of PoC and into production. Additionally, storage performance, read speed, needs to deliver performance to make sure data is always available to keep GPU resources busy and never idle waiting for data to arrive. This is why the All-Flash IBM ESS implementations are specified in the reference architecture.

Security and authentication start to become an issue. For example, many organizational data sources require authenticated access to their contents. IBM Spectrum Conductor provides specialized data connectors which act as templates for accessing external data sources as well as implements Kerberos authentication for secure connections to HDFS and other systems.

The exact architecture, vendors and components needed to build the network subsystem depend upon the organizational preference and skills. Note that thought needs to be put into a high-speed network subsystem that can serve the needs of the training cluster for storing the large datasets, then feeding the data to the compute servers in order to keep the GPUs highly utilized. A major requirement is to specify a network subsystem that delivers data at a constant rate, so the GPU resources never sit idle waiting for new data to arrive.

This is also the stage to implement **multitenancy** and develop a resource sharing plan across multiple data scientists opening up these powerful AI resources to a larger audience.



Stabilization & Production Configuration
IBM Recommended Infrastructure Systems
• IBM Power System AC922 (P9) and/or • IBM POWER S822LC (P8) for HPC and/or • IBM Elastic Storage Server (ESS) GS2/4S • High-speed network subsystem

IBM Recommended Solution Software
• IBM PowerAI Enterprose
• IBM Elastic Storage Server GUI
• IBM Spectrum Scale 5.x
• IBM Spectrum Scale RAID
• IBM Cloud Object Storage
• Red Hat Enterprise Linux for IBM Power Lille Endian

6.1. Scale-out and Expansion

In this stage of the AI project the same software stack used in the PoC and stabilization to production phase will also be used to scale out to serve a larger organization.

During the stabilization and production stage you were able to demonstrate the value derived from this AI environment and build demand within your organization for access to this environment. Some people will have legitimate AI projects, and some will want to run a PoC to educate themselves or convince themselves that an AI project will deliver value. This reference architecture, founded on IBM Spectrum Conductor, easily allows for the addition of new tenants (individual users, groups and lines of business) and through the use of templates easily creates a new AI environment. The resource sharing policies enable a tenant to have access to all available resources or be limited to a small subset to execute a PoC.

If not already, this is a good point in time for IT to becomes involved in the running and access control of this environment. Planning can start for connecting this environment to organizational management system. In addition to providing a better experience and accelerating the work of the data scientist, IBM Spectrum Conductor has also been designed as a systems management tool for distributed cluster systems. This administrative UI is suitable for operations type activities such as add users and groups, onboarding new frameworks and applications, adding new servers into the cluster and monitoring the health and utilization of the training cluster resources.

The multitenant and resource management capabilities also enable the operationalization of a software lifecycle to support the existing frameworks and applications as well as the rapid availability of new frameworks and versions.

6.1.1. Scaling, Availability and Reliability

As the size of the training cluster begins to grow and the AI function starts becoming a service that is expected to always be available, the roles of some of the infrastructure components need to be further defined. One activity in the growth of a training cluster is offloading function from the accelerated nodes and dedicating nodes to specific functions. Non-accelerated servers like the IBM Power Systems LC921 and LC922, as well as x86 based Linux systems, can be used to offload functions from the compute nodes.

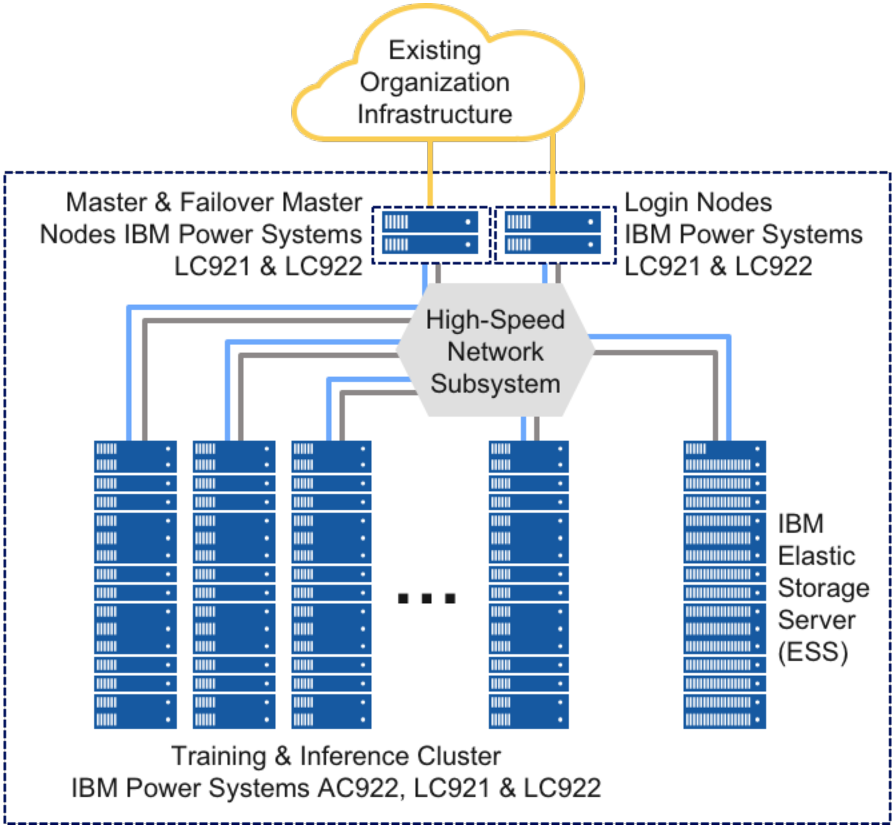
- **Machine learning and inferencing** nodes – Since ML and inferencing don’t always need to run on accelerated nodes, adding non-accelerated nodes will offload the compute nodes.
- **Master** node – Dedicating a node to be the IBM Spectrum Conductor master host. The master host function does not need to run on an accelerated system.

- **Failover Master** node – To provide higher availability to the AI environment, failover master hosts are either configured to use existing compute nodes or additional dedicated servers. A common configuration is two master hosts, one primary and failover are configured, and then a number of compute nodes are added to the failover list as backup to the failover hosts.
- **Login** nodes – To continue offloading services for compute nodes, a login server is added to the cluster. The login node function does not need to run on an accelerated system.

6.1.2. Network Considerations

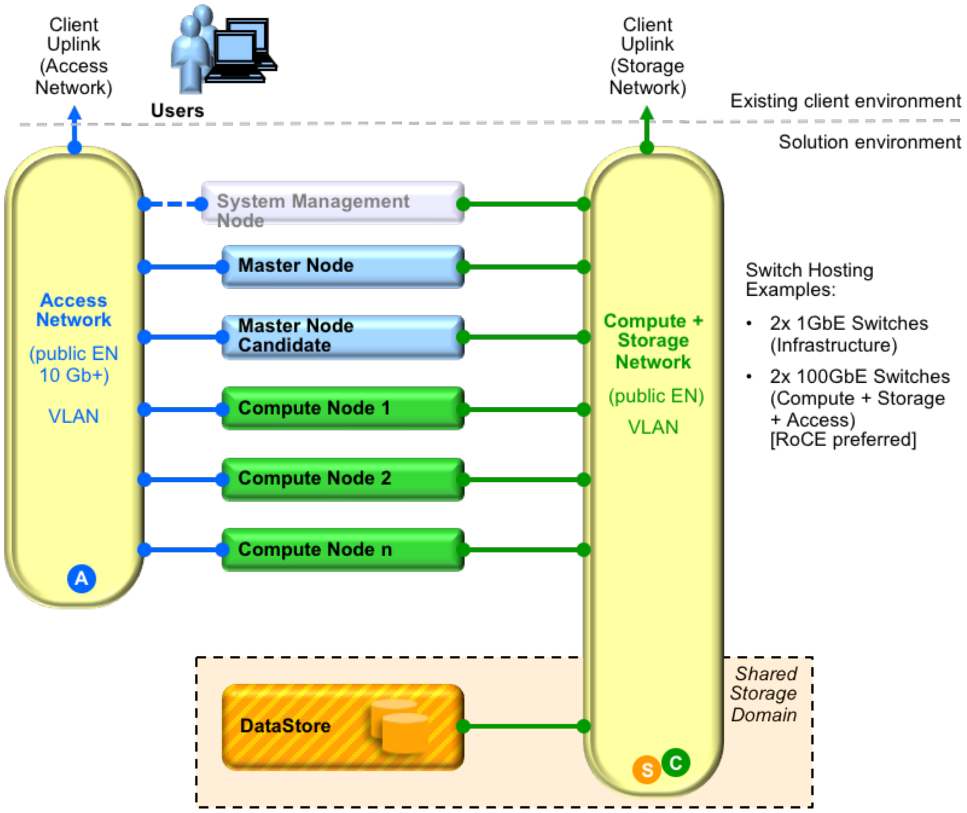
As the implementation grows creating different networks become needed to isolate traffic for:

- **Access** network connecting data scientists into the training cluster and providing access to the master node to interact with the IBM Spectrum Conductor Deep Learning Impact UI, As well as access to compute nodes and the storage system.
- **Storage** network, high speed connection from servers to storage system/ESS, to facilitate fast transfer of data while training. As well as North/South connections into the organizations other data sources and data lakes
- **Compute** network implemented with high speed connections between compute servers to facilitate inter process communications and weight synchronization and other activities needed when training distributed models.
- **Management** network (e.g., BMC) for direct control connection to all infrastructure elements. Also, a point of control for IT, which can either terminate in the training cluster or connect to the organization's larger management network

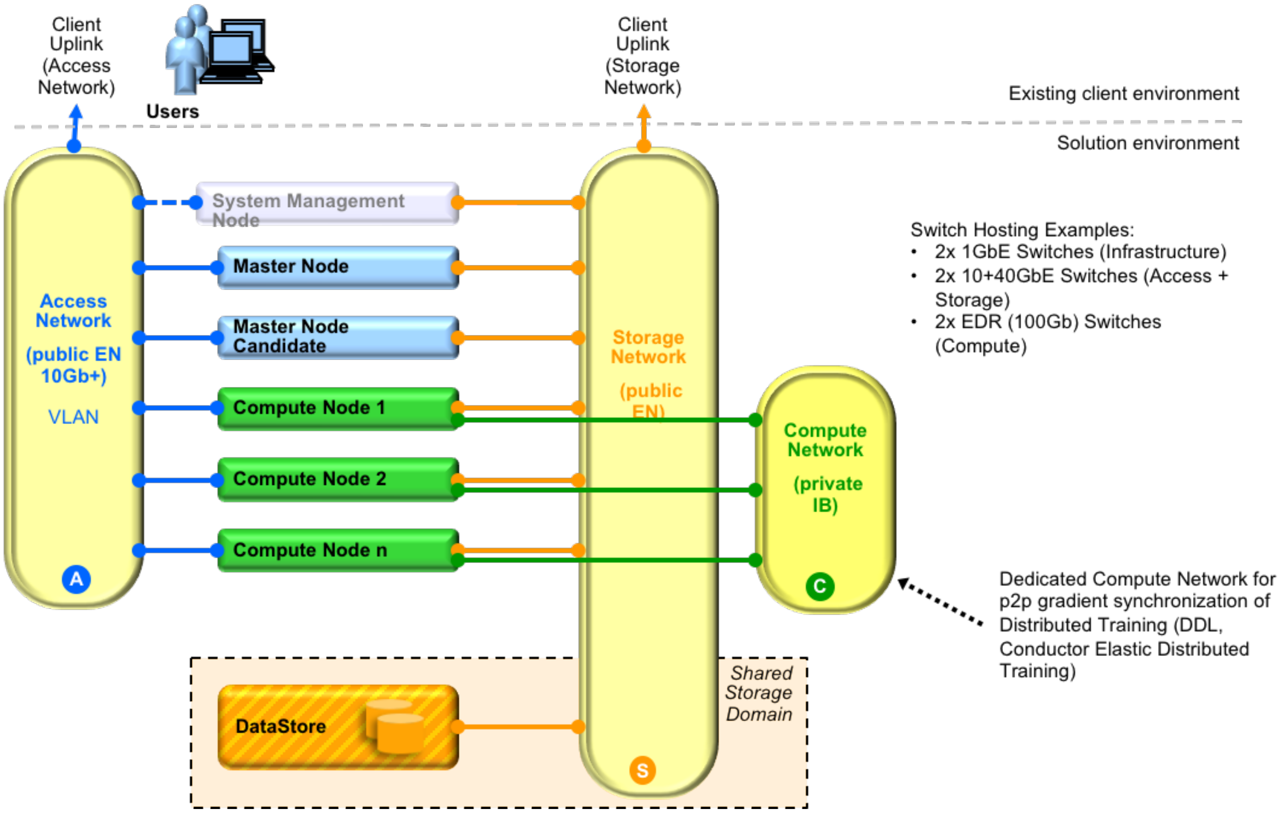

Existing Organization Infrastructure
Master & Failover Master Nodes IBM Power Systems LC921 & LC922
Login Nodes IBM Power Systems LC921 & LC922
High-Speed Network Subsystem
...
IBM Elastic Storage Server (ESS)
Training & Inference Cluster IBM Power Systems AC922, LC921 & LC922

7. Example Network Topologies

7.1. Ethernet



7.2. InfiniBand



Switch Hosting Examples:

- 2x 1GbE Switches (Infrastructure)
- 2x 10+40GbE Switches (Access + Storage)
- 2x EDR (100Gb) Switches (Compute)

Dedicated Compute Network for p2p gradient synchronization of Distributed Training (DDL, Conductor Elastic Distributed Training)

8. Example Bill of Materials

- Compute Node
- Login Node
- ESS Management Node
- ESS IO Node

8.1. Compute Node

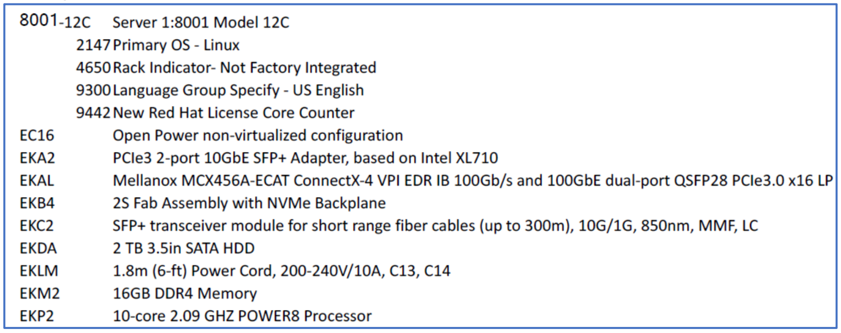
- IBM Power System AC922 (8335-GTG)

8335-GTG	Server 1:8335 Model GTG	1
2147	Primary OS - Linux	1
4650	Rack Indicator- Not Factory Integrated	1
9300	Language Group Specify - US English	1
9442	New Red Hat License Core Counter	40
5765-AIE 0002	IBM PowerAI Enterprise V1 per Virtual Server	1
5773-AIE 1849	3-Year SWMA for 5765-AIE per Virtual Server 24x7 Support	1
5773-AIE 1850	3-Year SWMA for 5765-AIE per Virtual Server SW Maint Registration	1
EB2X	AC Power Supply - 2200 WATT (227V)	2
EC4J	Air-Cooled 16GB SXM2 FF 300W NVIDIA Volta GPU	4
EC64	PCIe4.04 LP 2-port 100Gb EDR IB CAPI adapter	1
EJTY	Rack-mount Slide Rail Kit	1
ELU5	1.92 TB 2.5in SATA/SSD Disk Drive	2
EM63	32 GB DDR4 2666 RDIMM	16
EN0T	PCIe2 LP 4-Port (10Gb+1GbE) SR+RJ45 Adapter	2
EP0M	20-core 2.0 GHz (2.87 GHz Turbo) POWER9	2
EPAM	Power Cord 4.3 M (14.10-foot), Drawer to IBM PDU, 250V/16A	2
ERBZ	No Bulk Packaging Specify	1
ESC5	S&H-a	1

- Knowledge Center: https://www.ibm.com/support/knowledgecenter/en/POWER9/p9hdx/8335_gtg_landing.htm

8.2. Login Node

- IBM Power System S821LC (8001-12C)



8001-12C	Server 1:8001 Model 12C
2147	Primary OS - Linux
4650	Rack Indicator- Not Factory Integrated
9300	Language Group Specify - US English
9442	New Red Hat License Core Counter
EC16	Open Power non-virtualized configuration
EKA2	PCIe3 2-port 10GbE SFP+ Adapter, based on Intel XL710
EKAL	Mellanox MCX456A-ECAT ConnectX-4 VPI EDR IB 100Gb/s and 100GbE dual-port QSFP28 PCIe3.0 x16 LP
EKB4	2S Fab Assembly with NVMe Backplane
EKC2	SFP+ transceiver module for short range fiber cables (up to 300m), 10G/1G, 850nm, MMF, LC
EKDA	2 TB 3.5in SATA HDD
EKLM	1.8m (6-ft) Power Cord, 200-240V/10A, C13, C14
EKM2	16GB DDR4 Memory
EKP2	10-core 2.09 GHZ POWER8 Processor

- Knowledge Center: https://www.ibm.com/support/knowledgecenter/en/POWER8/p8hdx/8001_12c_landing.htm

8.3. IBM ESS Management Server

- IBM Elastic Storage Server (ESS) Management Server (5148-21L)

Part Number	Description	Quantity
5148-21L	ESS Management Server 1:5148 Model 21L 1	1
5771	SATA Slimline DVD-RAM Drive 1	1
6577	Power Cable - Drawer to IBM PDU, 200-240V/10A 2	2
EB5A	3M EDR IB Optical Cable QSFP28 4	4
EC3E	PCIe3 LP 2-port 100Gb EDR IB Adapter x16 2	2
EJTT	Front Bezel for 12-Bay BackPlane 1	1
EL1A	AC Power Supply - 900W 2	2
EL3T	Storage Backplane 12 SFF-3 Bays/DVD Bay 1	1
EL3X	PCIe3 LP 2-port 10GbE NIC&RoCE SFP+ Copper Adapter 1	1
EL4M	PCIe2 LP 4-port 1GbE Adapter 1	1
ELD5	600GB 10K RPM SAS SFF-3 Disk Drive (Linux) 2	2
ELPD	10-core 3.42 GHz POWER8 Processor Card 1	1
EM98	64 B DDR4 Memory 4	4

8.3.1. IBM ESS Options

ESS Option - Storage Capacity					
GS1S	96	-	367	TB	Flash
GS2S	192	-	734	TB	Flash
GS4S	381	-	1468	TB	Flash
GL1S	328	-	820	TB	HDD
GL2S	.68	-	1.7	PB	HDD
GL4S	1.3	-	3.3	PB	HDD
GL6S	2	-	5.1	PB	HDD
GL4C	4.2	-	4.2	PB	HDD
GH14	3.84 1428	- -	15.36 3708	TB TB	Flash HDD
GH24	3.84 1520	- -	15.36 4077	TB TB	Flash HDD

- Knowledge Center: https://www.ibm.com/support/knowledgecenter/en/5148-21L/p8hdx/5148_21l_landing.htm

8.4. IBM Elastic Storage Server IO Server

- IBM Elastic Storage Server (ESS) IO/Data Server (5148-22L)

Part Number	Description	Quantity
5148-22L	5148 Model 22L	1
5771	SATA Slimline DVD-RAM Drive	1
6577	Power Cable - Drawer to IBM PDU, 200-240V/10A	2
EB5A	3M EDR IB Optical Cable QSFP28	4
EC3E	PCIe3 LP 2-port 100Gb EDR IB Adapter x16	2
ECE3	3.0M SAS AA12 Cable (Adapter to Adapter)	4
EL1B	AC Power Supply - 1400W (200-240 VAC)	2
EL3W	2U SAS RAID 0,5,6,10 Controller + Back plane	1
EL3X	PCIe3 LP 2-port 10GbE NIC&RoCE SFP+ Copper Adapter	1
EL4M	PCIe2 LP 4-port 1GbE Adapter	1
ELD5	600GB 10K RPM SAS SFF-3 Disk Drive (Linux)	
ELPD	10-core 3.42 GHz POWER8 Processor Card	2
EM98	64 GB DDR4 Memory	4
EN02	3m (9.8-ft), 10Gb E'Net Cable SFP+ Act Twinax Copper	2
ESA5	LSI SAS Controller 9305-16E 12GB/S host bus adapter	4

- Knowledge Center: https://www.ibm.com/support/knowledgecenter/en/POWER8/p8hdx/5148_22l_landing.htm

8.5. Network Switches

8.5.1. 10 GbE - IBM G8052 (7120-48E) Switch

- Sourced from Lenovo

Part Number	Description	Quantity
7120-48E	Switch 1:7120 Model 48E	1
4650	Rack Indicator- Not Factory Integrated	1
6458	Power Cord 4.3m (14-ft), Drawer to IBM PDU (250V/10A)	2
EDT2	1U AIR DUCT MED	1
ESC7	S&H	1
EU27	IBM System Networking Adjustable 19inch 4 Post Rail Kit	1

8.5.2. InfiniBand - IBM 8828-E36 (SB7700-ES2F) EDR Switch

- Made by Mellanox

Part Number	Description	Quantity
8828-E36	Switch 1:8828 Model E36	1
4650	Rack Indicator- Not Factory Integrated	1
EB52	2.0M EDR IB Copper Cable QSFP28	8
EB54	1.5M EDR IB Copper Cable QSFP28	4
EB5C	10M EDR IB Optical Cable QSFP28	5
EB5D	15M EDR IB Optical Cable QSFP28	4

8.5.3. 10 GbE - 7120-24E (G8124E) RackSwitch

- Sourced from Lenovo Model 24E

Part Number	Description	Quantity
7120-24E	10GbE Switch:7120 Model 24E	1
4650	Rack Indicator- Not Factory Integrated	1
6458	Power Cord 4.3m (14-ft), Drawer to IBM PDU (250V/10A)	2
EDT1	1U Air Duct Long	1
ESC7	S&H	1
EU27	IBM System Networking Adjustable 19inch 4 Post Rail Kit	1

9. References

- IBM **PowerAI**: Deep Learning Unleashed on IBM Power Systems Servers, including Chapter 6. Introduction to IBM **Spectrum Conductor Deep Learning Impact**, http://www.redbooks.ibm.com/abstracts/sg248409.html
- IBM **Spectrum Conducto**r and IBM Spectrum Conductor with Spark, http://www.redbooks.ibm.com/abstracts/redp5379.html
- IBM **Spectrum Scale** (formerly GPFS), http://www.redbooks.ibm.com/abstracts/sg248254.html
- Introduction Guide to the IBM **Elastic Storage Server**, http://www.redbooks.ibm.com/abstracts/redp5253.html
- Monitoring Overview for IBM **Spectrum Scale** and IBM **Elastic Storage Server**, http://www.redbooks.ibm.com/abstracts/redp5418.html
- IBM Power System **AC922** Introduction and Technical Overview, https://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/redp5472.html
- IBM Power System **S822LC** Technical Overview and Introduction, http://www.redbooks.ibm.com/abstracts/redp5283.html
- IBM Power Systems **S812L** and **S822L** Technical Overview and Introduction, http://www.redbooks.ibm.com/abstracts/redp5098.html

10. Notices

IBM Systems
3039 Cornwallis Road
RTP, NC 27709

Produced in the United States of America

References in the publication to IBM products or services do not imply that IBM intends to make them available in all countries in the IBM operates.

♻ Please Recycle